

LAREX - Ein Werkzeug zur Layout-Analyse und Segmentierung von frühen Buchdrucken

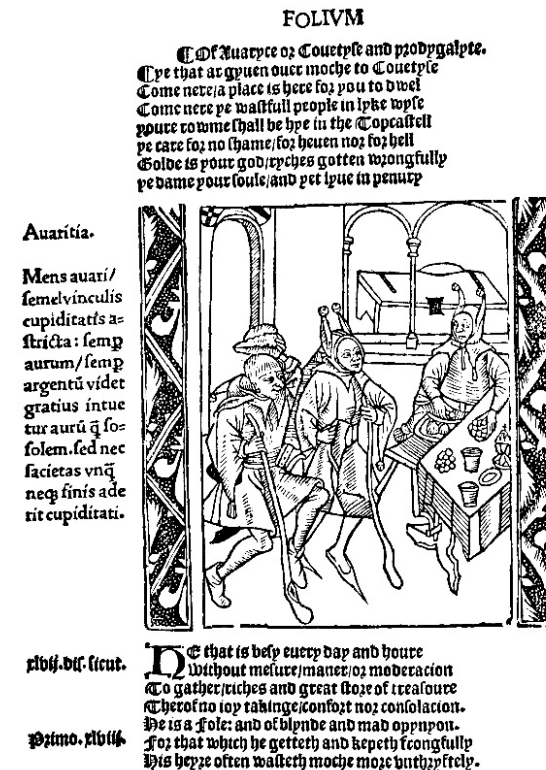
Christian Reul

Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik
Universität Würzburg

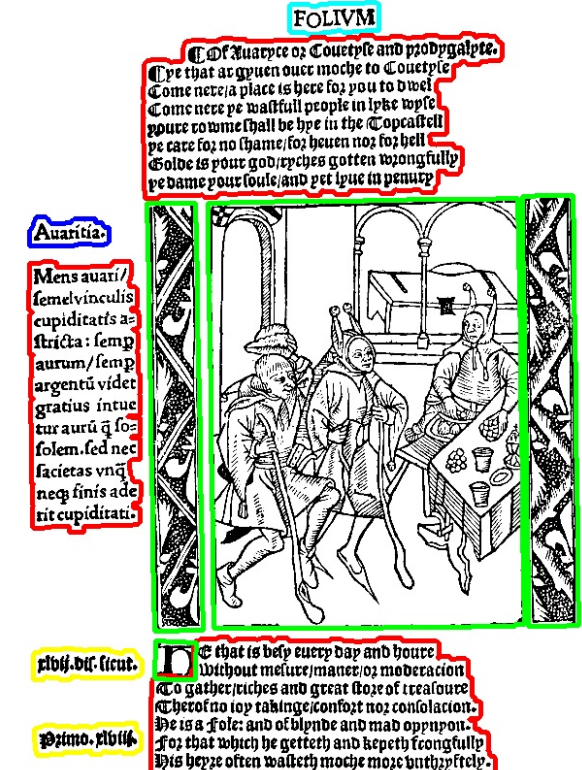
17.03.2017

Motivation und Ansatz

- Segmentierung vor OCR
meist unerlässlich.
- Text/Bild-Trennung oft
nicht ausreichend.
- Häufig zusätzliche semantische
Auszeichnung erwünscht.
- Semi-automatisch.
- Intuitiv, anpassbar und nachvollziehbar.
- Open Source.
- Layout Analysis and Region Extraction.



Input



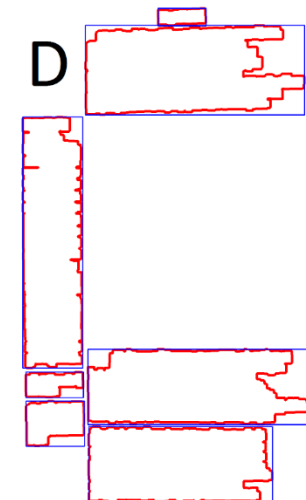
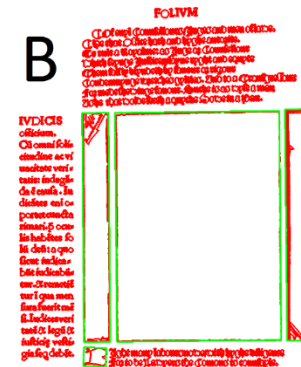
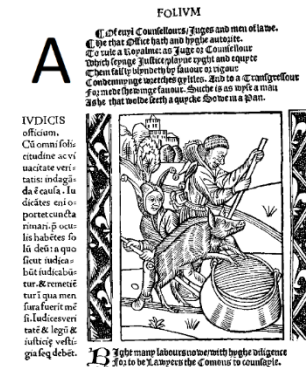
Beispiel-Output

Workflow

Input: Scan einer Seite.

Output: Klassifizierte Segmente als PageXML.

1. Pre-processing (A).
2. Bilddetektion (A → B).
 1. Optionale Dilatation des Vordergrundes.
 2. Bild/Nicht-Bild-Klassifikation.
3. Entfernen der Bilder.
 1. Dilatation des Vordergrundes.
 2. Regelbasierte Klassifikation.
3. Grobe Textklassifikation (C → D, E).
 1. Dilatation des Vordergrundes.
 2. Regelbasierte Klassifikation.
4. Manuelle Korrektur und feine Textklassifikation (E → F).
5. Konvertierung in Ausgabeformat.



Globale Optimierung - Überblick

- Einzelne Bücher meist homogen.
 - Abstände.
 - Positionen (z. B. von Marginalien).
 - ...
- Idee: Finden einer passenden, individuellen Layout-Maske.
- Manuelle Optimierung durch
 - Anpassung der Dilatations-Parameter.
 - Hinzufügen und Verschieben von Regionen-Typen und deren Positionen.
 - Anpassung der Regionen-Mindestgröße.
 - ...

III
mufedogmata paffim feminarente te aiebam amicu qui fite
Ib Nafum apud noftrates faceres: Itaq in verba fidetio fte
Chusre adhorari non definam: vt in his noftris lucubrati
his fauorabilis affenfus: curamq pntem adhibeat. No eni
dubito quin erallus quida auribus: corde obfcurato: impex
is fup pelis / feilili pallia / trofufuliq: centuculo amicti: ma
nus fanguinarias temere nobis impingant: horum glutino
fatum faucium larratus larualeq: ac toruofas exclamatio
nes: celestinaq: nauticu/cauillum exoticum: fellularia (q: ca
ftigationes: te duce: te pfgide: te quoq: tutore facile fupera
bimus. Eit etiam nobis altior cothurnus: funt nobis fyrra
ta longiora: dexter quoq: Apollo cum iucida camerarum
chorea: nodu: aedes ac noftra diuerforia defertur. calli ftiqui
dem manu palladis artes: facraq: Mithigeos aramithureva
pido/frequentiq: libo veneramur. Igitur dulciffime fautori
ventis impera: & nauticulis noftris ac phafelis: falubres au
ras precare. Vale. Datum Eriburgi. Kal. Februariis. Anno
domini. M.CCCC.XCVII.

CARMEN eiusdem d. St. Brant.

S I nuch nunc ppari sacros concederet arcus:
Verteret in nitidos & mea verba pedes:
Et daret arguti cultiffima plectra leporis:
Ad te migraret/hoc duce/culta lyra:
Nil rude: nil repidum: venerando nomine dignu
Esse tuo poterit: tu quia dulce canis.
Siue voles numeris facundos neclere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicrone loqui.
Ia modo plaudetis: tuis foelix Germania nymphis:
Quas fontes Rheni/Danubiique fouent.
Non fumus auersi a musis/& Apolline dextros:
Tangere iam didicit Theutona terra lyram.
Mufica noftrates /phebo duce/venit ad bras.
Scimus & argutos voce fonare modos:
A regione precul noftra: permiffidos vnda
Sacra feater: nec nos Aona praeta figant:

Carmen ad
S. Brant.

Quare Ger
mani rardius
ad mufas ve
nerunt come
tium.

Segmentierungsergebnis
mit Default Setup

III
mufedogmata paffim feminarente te aiebam amicu qui fite
Ib Nafum apud noftrates faceres: Itaq in verba fidetio fte
Chusre adhorari non definam: vt in his noftris lucubrati
his fauorabilis affenfus: curamq pntem adhibeat. No eni
dubito quin erallus quida auribus: corde obfcurato: impex
is fup pelis / feilili pallia / trofufuliq: centuculo amicti: ma
nus fanguinarias temere nobis impingant: horum glutino
fatum faucium larratus larualeq: ac toruofas exclamatio
nes: celestinaq: nauticu/cauillum exoticum: fellularia (q: ca
ftigationes: te duce: te pfgide: te quoq: tutore facile fupera
bimus. Eit etiam nobis altior cothurnus: funt nobis fyrra
ta longiora: dexter quoq: Apollo cum iucida camerarum
chorea: nodu: aedes ac noftra diuerforia defertur. calli ftiqui
dem manu palladis artes: facraq: Mithigeos aramithureva
pido/frequentiq: libo veneramur. Igitur dulciffime fautori
ventis impera: & nauticulis noftris ac phafelis: falubres au
ras precare. Vale. Datum Eriburgi. Kal. Februariis. Anno
domini. M.CCCC.XCVII.

CARMEN eiusdem d. St. Brant.

S I nuch nunc ppari sacros concederet arcus:
Verteret in nitidos & mea verba pedes:
Et daret arguti cultiffima plectra leporis:
Ad te migraret/hoc duce/culta lyra:
Nil rude: nil repidum: venerando nomine dignu
Esse tuo poterit: tu quia dulce canis.
Siue voles numeris facundos neclere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicrone loqui.
Ia modo plaudetis: tuis foelix Germania nymphis:
Quas fontes Rheni/Danubiique fouent.
Non fumus auersi a musis/& Apolline dextros:
Tangere iam didicit Theutona terra lyram.
Mufica noftrates /phebo duce/venit ad bras.
Scimus & argutos voce fonare modos:
A regione precul noftra: permiffidos vnda
Sacra feater: nec nos Aona praeta figant:

Carmen ad
S. Brant.

Quare Ger
mani rardius
ad mufas ve
nerunt come
tium.

Default-Positionen für Seitenzahl (cyan)
und Marginalien (gelb).

Globale Optimierung - Dilatation

- Erweiterung von Vordergrundpixeln um X/Y Pixel in x/y-Richtung.
- Gut passender Wert essentiell:
 - Zu niedrig: Trennung zusammen gehöriger Zeichen/Wörter/Zeilen.
 - Zu hoch: Verschmelzung nicht zusammen gehöriger Blöcke.
- Optimierung im Normalfall sehr einfach.

III

muse dogmata passim seminarent: te aiebam amicu qui stit-
lo Nasum apud nostrates faceres: Itaq; in verba fidetior fa-
ctus: te adhortari non desinam: vt in his nostris lucubrati-
bus fauorabiles assensus: curamq; pntem adhibeas. No eni
dubito: quin crassis quida auribus: corde obstinato: impex-
is suppelis/scissili pallia stro: futuliq; centuculo amicti: ma-
nus sanguinarias temere nobis impingant. horum glutino-
satum faucium latratus larualeq; ac tortuosas exclamatio-
nes: celeulmaq; nauticu/cauillum exoticum: sellulariaq; ca-
stigationes: te duce: te preside. te quoq; tutore facile superas-
bimus. Est etiam nobis altior cothurnus: sunt nobis syrma-
ta longiora: dexter quoq; Apollo cum iucunda camænarum
chorea: nodu ædes ac nostra diuersoria deseruit. casta liqui-
dem manu palladis artes. sacraq; Mathæseos aram: thure va-
pido/frequentiq; libo veneramur. Igitur dulcissime fautori:
ventis impera: & nauticulis nostris ac phaselis: salubres aus-
ras precare. Vale. Datum Friburgi. Kalē. Februariis. Anno
domini. M.CCCC.XCVII.

CARMEN eiusdem: ad Se. Brant.

S mihi nunc pæn sacros concederet arcus:
Verteret in nitidos & mea verba pedes:
Et daret arguti cultissima plectra leporis:
Ad te migraret/hoc duce/culta lyra:
Nil rude: nil tepidum: venerando nomine dignū
Esse tuo poterit: tu quia dulce canis.
Sue voles numeris facundosnectere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicerone loqui.
Iā modo plaudet: tuis foelix Germania nymphis:
Quas fontes Rheni/Danubiiq; fount.
Non sumus auersi a musis/& Apolline dextro:
Tangere iam didicit Theutona terra lyram.
Musica nostrates/phebo duce/venit ad oras.
Scimus & argutos voce sonare modos.
A regione procul nostra: permissidos vnda
Sacræ scater: nec nos Aona prata rigant:
a.iii.

Carmen ad
S. Brant.

Quare Ger-
mani rardius
ad musas ve-
nerint cōmer-
tium.

Default Parameter

III

muse dogmata passim seminarent: te aiebam amicu qui stit-
lo Nasum apud nostrates faceres: Itaq; in verba fidetior fa-
ctus: te adhortari non desinam: vt in his nostris lucubrati-
bus fauorabiles assensus: curamq; pntem adhibeas. No eni
dubito: quin crassis quida auribus: corde obstinato: impex-
is suppelis/scissili pallia stro: futuliq; centuculo amicti: ma-
nus sanguinarias temere nobis impingant. horum glutino-
satum faucium latratus larualeq; ac tortuosas exclamatio-
nes: celeulmaq; nauticu/cauillum exoticum: sellulariaq; ca-
stigationes: te duce: te preside. te quoq; tutore facile superas-
bimus. Est etiam nobis altior cothurnus: sunt nobis syrma-
ta longiora: dexter quoq; Apollo cum iucunda camænarum
chorea: nodu ædes ac nostra diuersoria deseruit. casta liqui-
dem manu palladis artes. sacraq; Mathæseos aram: thure va-
pido/frequentiq; libo veneramur. Igitur dulcissime fautori:
ventis impera: & nauticulis nostris ac phaselis: salubres aus-
ras precare. Vale. Datum Friburgi. Kalē. Februariis. Anno
domini. M.CCCC.XCVII.

CARMEN eiusdem: ad Se. Brant.

S mihi nunc pæn sacros concederet arcus:
Verteret in nitidos & mea verba pedes:
Et daret arguti cultissima plectra leporis:
Ad te migraret/hoc duce/culta lyra:
Nil rude: nil tepidum: venerando nomine dignū
Esse tuo poterit: tu quia dulce canis.
Sue voles numeris facundosnectere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicerone loqui.
Iā modo plaudet: tuis foelix Germania nymphis:
Quas fontes Rheni/Danubiiq; fount.
Non sumus auersi a musis/& Apolline dextro:
Tangere iam didicit Theutona terra lyram.
Musica nostrates/phebo duce/venit ad oras.
Scimus & argutos voce sonare modos.
A regione procul nostra: permissidos vnda
Sacræ scater: nec nos Aona prata rigant:
a.iii.

Carmen ad
S. Brant.

Quare Ger-
mani rardius
ad musas ve-
nerint cōmer-
tium.

Optimierte Parameter

Globale Optimierung - Positionen

- Zuweisung nur bei vollständiger Überdeckung eines Blocks.
- Beliebige Anzahl von Positionen.
- Komfortables Hinzufügen, Löschen und Verschieben möglich.

III
muse dogmata pallum seminarent: te aiebam amicum qui stilo
Nasum apud nostrates faceres: Itaque in verba fideior fa-
ctus: te adhortari non desinam: vt in his nostris lucubrato-
bus fauorabiles assensus: curamque priorem adhibeas. Nō eni
dubito: quin et assis quidā auribus: corde obstinato: impex-
is suppelis/ scilicet pallia/ tro: futulique: centuculo amicti: ma-
nus sanguinarias temere nobis impingant. horum glutino-
farum faucium lacratu larualeque ac tortuosas exclamatio-
nes: celestinaque nautici/ cauillum exoticum: scellulariaque
stigationes: te duce: te p̄sede. te quoque tutore facile superas-
himus. Est etiam nobis altior cothurnus: sunt nobis lyra
et longiora: dexter quoque Apollo cum iucunda camœnarum
chorea: nōdū ades ac nostra diuersoria deferuit. casta siqui-
dem manu palladis artes. sacraque Mathæos aram: thure va-
pido/ frequentique libo veneramur. Igitur dulcissime fautor:
ventis impera: & nauiculis nostris ac phaselis: salubres au-
ras precare. Vale. Datum Friburgi. Kalē. Februariis. Anno
domini. M. CCCC. XC VII.

CARMEN eiusdem ad S. Brant.

Sil mihi nunc pan sacros concederet arcus:
Vteret in nitidos & mea verba pedes:
Et daret arguti cultissima plectra leporis:
Ad te migraret/ hoc duce/ culta lyra.
Nil rudenil tepidum: venerando nomine dignū
Esse tuo poterit: tu quia dulce canis.
Sive voles numeris facundos neclere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicerone loqui.
Ita modo plaude: tuis felix Germania nymphis:
Quas fontes Rheni/ Danubiique fount.
Non sumus auersi a musis/ & Apolline dextros
Tangere iam didicit Theutona terra lyram.
Mufica nostrates/ p̄hebo duce/ venit ad oras.
Scimus & argutos voce sonare modos.
A regione procul nostra: permessidos vnda
Sacta featur: nec nos Aona prata rigant.

a. iii.

Carmen ad
S. Brant.

Quare Ger-
mani rardius
ad musas ve-
nerint cōme-
tium.

Default Positionen

III
muse dogmata pallum seminarent: te aiebam amicum qui stilo
Nasum apud nostrates faceres: Itaque in verba fideior fa-
ctus: te adhortari non desinam: vt in his nostris lucubrato-
bus fauorabiles assensus: curamque priorem adhibeas. Nō eni
dubito: quin et assis quidā auribus: corde obstinato: impex-
is suppelis/ scilicet pallia/ tro: futulique: centuculo amicti: ma-
nus sanguinarias temere nobis impingant. horum glutino-
farum faucium lacratu larualeque ac tortuosas exclamatio-
nes: celestinaque nautici/ cauillum exoticum: scellulariaque
stigationes: te duce: te p̄sede. te quoque tutore facile superas-
himus. Est etiam nobis altior cothurnus: sunt nobis lyra
et longiora: dexter quoque Apollo cum iucunda camœnarum
chorea: nōdū ades ac nostra diuersoria deferuit. casta siqui-
dem manu palladis artes. sacraque Mathæos aram: thure va-
pido/ frequentique libo veneramur. Igitur dulcissime fautor:
ventis impera: & nauiculis nostris ac phaselis: salubres au-
ras precare. Vale. Datum Friburgi. Kalē. Februariis. Anno
domini. M. CCCC. XC VII.

CARMEN eiusdem ad S. Brant.

Sil mihi nunc pan sacros concederet arcus:
Vteret in nitidos & mea verba pedes:
Et daret arguti cultissima plectra leporis:
Ad te migraret/ hoc duce/ culta lyra.
Nil rudenil tepidum: venerando nomine dignū
Esse tuo poterit: tu quia dulce canis.
Sive voles numeris facundos neclere sensus:
Seu fidibus lyricis: optimus arte vales.
Expertes uumeris seu malis condere voces:
Audes magnifico cum Cicerone loqui.
Ita modo plaude: tuis felix Germania nymphis:
Quas fontes Rheni/ Danubiique fount.
Non sumus auersi a musis/ & Apolline dextros
Tangere iam didicit Theutona terra lyram.
Mufica nostrates/ p̄hebo duce/ venit ad oras.
Scimus & argutos voce sonare modos.
A regione procul nostra: permessidos vnda
Sacta featur: nec nos Aona prata rigant.

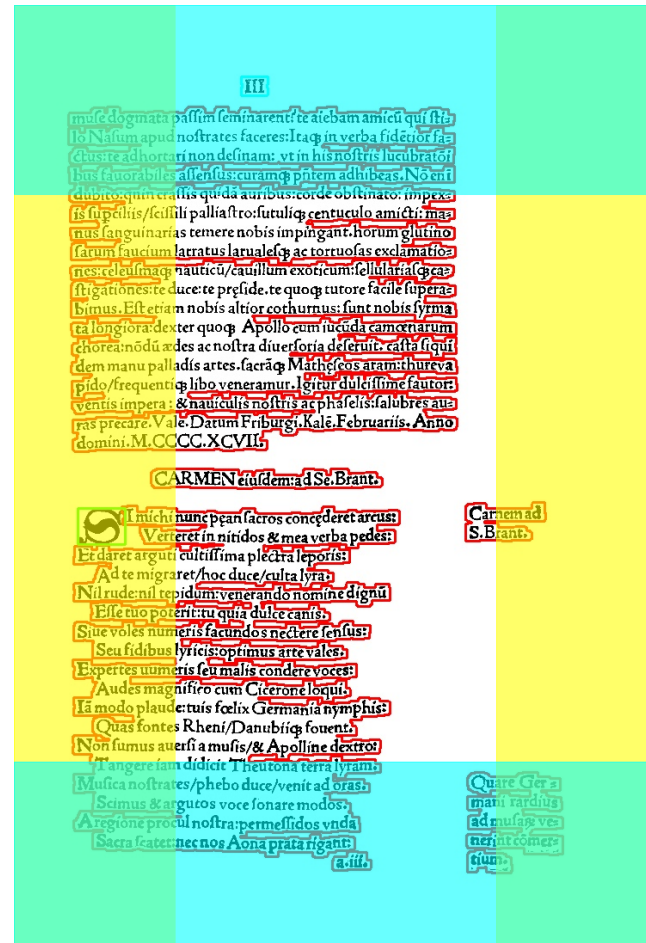
a. iii.

Carmen ad
S. Brant.

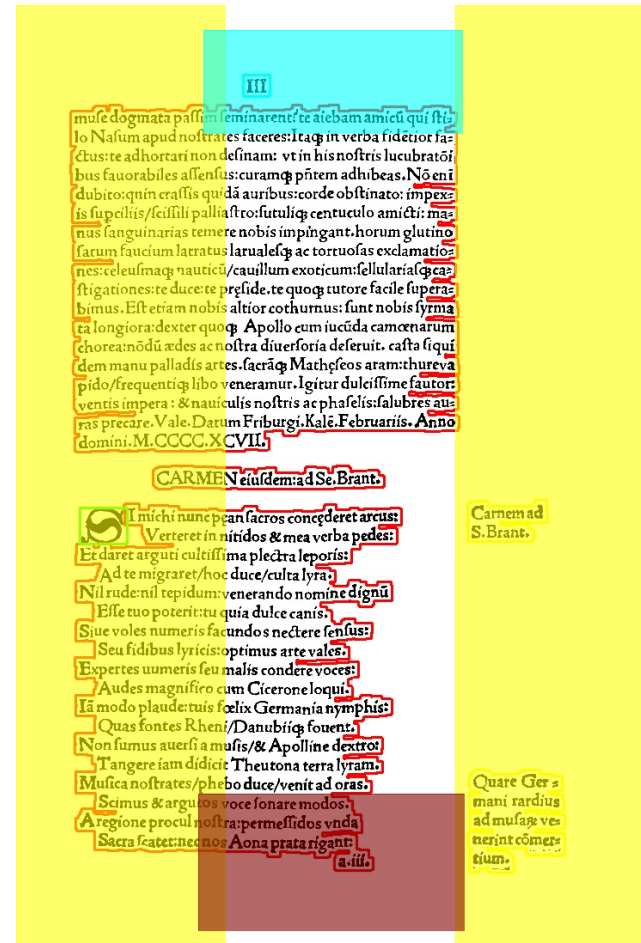
Quare Ger-
mani rardius
ad musas ve-
nerint cōme-
tium.

Optimierte Positionen

Globale Optimierung - Ergebnis



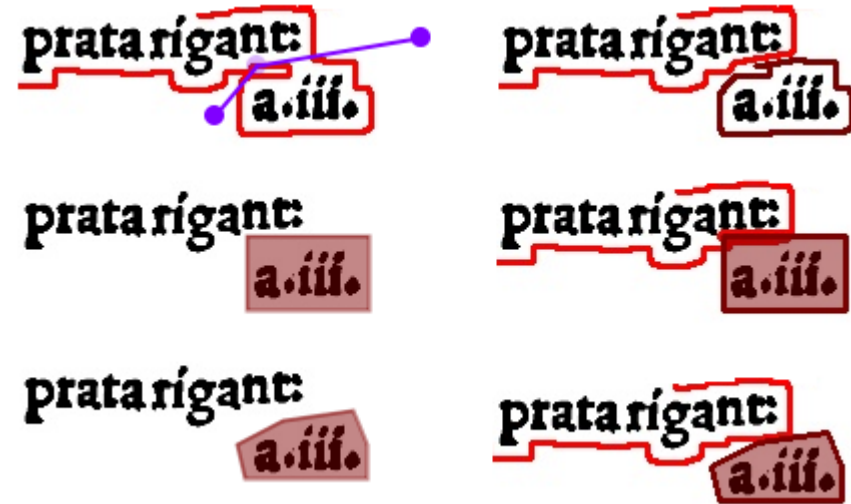
Default Setup und Ergebnis



Optimiertes Setup und Ergebnis

Lokale, manuelle Korrekturen

- Löschen von Regionen.
- Ändern des Typs.
- Blocktrennung durch Linien.
- Manuelle Auszeichnung von Regionen durch Rechtecke oder Polygone.



Feinsegmentierung auf Zeilenebene

- Nutzung der Tesseract Zeilensegmentierung.
- Anschließend Blocktrennung anhand der Zeilen möglich.
- Zeitintensiv (Sammeln!).

De inutilibus libris.
Inter precipuos pars est mihi reddita stultos
Prima:rego docili vastaꝓ vela manu.
En ego possideo multos:quos raro libellos
Perlego:tum lectos negligo:nec sapio.

De inutilibus libris.
Inter precipuos pars est mihi reddita stultos
Prima:rego docili vastaꝓ vela manu.
En ego possideo multos:quos raro libellos
Perlego:tum lectos negligo:nec sapio.

De inutilibus libris.
Inter precipuos pars est mihi reddita stultos
Prima:rego docili vastaꝓ vela manu.
En ego possideo multos:quos raro libellos
Perlego:tum lectos negligo:nec sapio.

De inutilibus libris.
Inter precipuos pars est mihi reddita stultos
Prima:rego docili vastaꝓ vela manu.
En ego possideo multos:quos raro libellos
Perlego:tum lectos negligo:nec sapio.

Blocksegmentierung (ol), Zeilensegmentierung (or),
markierte Überschrift (ul), Endergebnis (ur).

Petrarcha
P. Beroal-
dus.
ci calam argutia: egyptiꝓ papyri crassitudo, no-
stris fudoribus uix laris faceret. Nostre itaq; rudi-
raris exercitamenta, hilari fronte iucundissimoꝓ
nietis examine trutinabis. Laudabis discipuli tui
audaciam: qui sui preceptoris saluberrima rhyth-
mara: lariali (quauis dura ac bulbutienti) lingua e-
rheutonico liguagio uertit. Nec id pudoris loco
habendū duco. cū & Frācisci Petrarche philoso-
phi Stoici: ac Meonii uaris seclatoris celeberrī: cā-
rões ulgares: uernaculaꝓ dictamia, Philippus Be-
roaldus Bononiē, eque pceptor meus, omniꝓ
uerustatis candidissimus in quistor (cui etiam iam
dudū primas partes in oi dicendi genere Italia: fel-
sinaꝓ uirum tribuit) in latinū sermonē uertere nō
est dedignatus. Idē Boccatil interpres esse uoluit.
Maximi igit spectatissimꝓ honoris pmiꝓ loco
duximꝓ: q; me sermonis tui uernaculi (quo egregie
polles) interprē esse uoluit. Illud nepe inter preci-
pua humanitatis gēnera locari solet: cū a pcepto-
re discipulus ad honores egregios eleuat: Tātū igit
tur te facio: q;ti Iocrates Platoꝓ Socrate, sanctissi-
mū fecere. Nec cedo ea ire Theophrasto: q; Aristo-
tilē peripatericoꝓ fundatore maximū maxi fecit.
Ecq;d priscos cito: cū satis supꝓ mea i te obserua-
bilꝓ piꝓq; ueneratio, lōge lateꝓ p fines Germaniꝓ
eruditissimāꝓ latī scholas, me preconē, itonuī.
testis ē mihi Hubertinꝓ Clericus Cresserinas: mar-
chionis Montiserrati & pedemōranꝓ poeta stipe-
darius: apud quē Cassale tuā egregiam uirtutem
expolui. Interrogabat enī uir iste apprimē doctus:

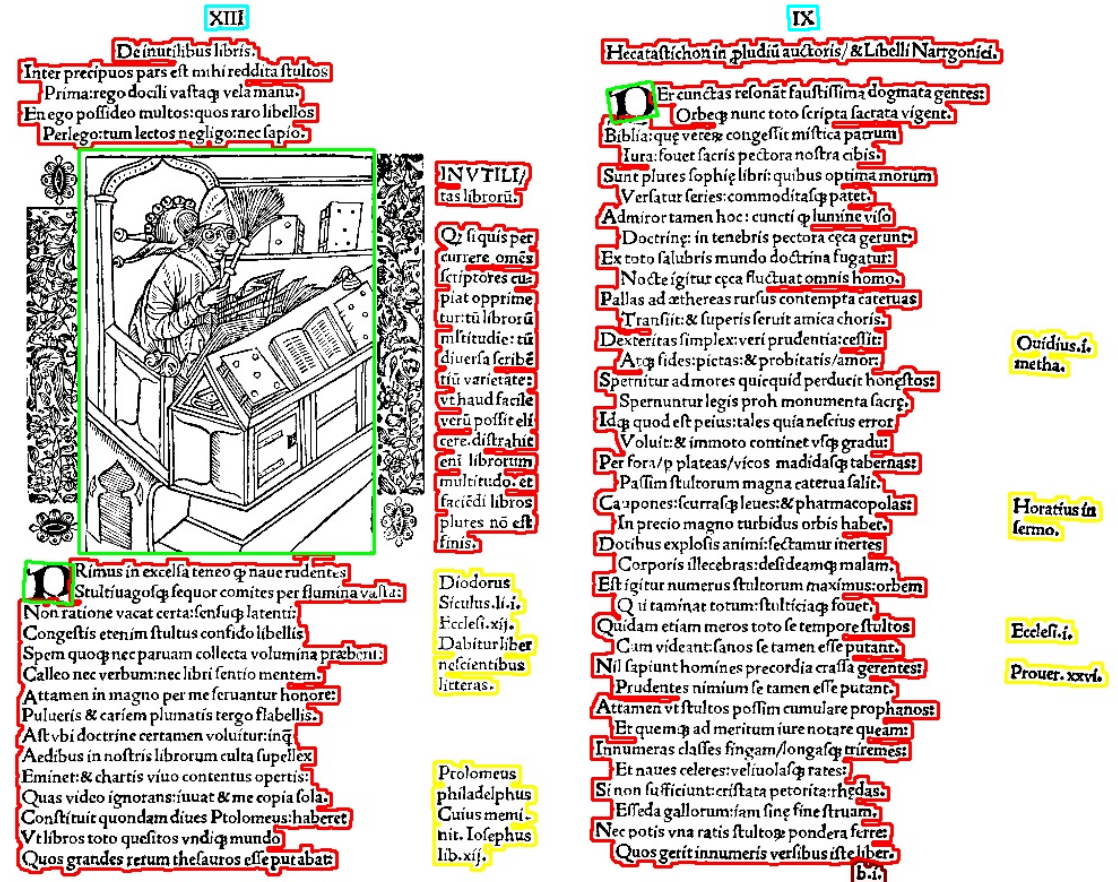
Ergebnis Blocksegmentierung

Petrarcha
P. Beroal-
dus.
ci calam argutia: egyptiꝓ papyri crassitudo, no-
stris ludoribus uix laris faceret. Nostre itaq; rudi-
raris exercitamenta, hilari fronte iucundissimoꝓ
nietis examine trutinabis. Laudabis discipuli tui
audaciam: qui sui preceptoris saluberrima rhyth-
mara: lariali (quauis dura ac bulbutienti) lingua e-
rheutonico liguagio uertit. Nec id pudoris loco
habendū duco. cū & Frācisci Petrarche philoso-
phi Stoici: ac Meonii uaris seclatoris celeberrī: cā-
rões ulgares: uernaculaꝓ dictamia, Philippus Be-
roaldus Bononiē, eque pceptor meus, omniꝓ
uerustatis candidissimus in quistor (cui etiam iam
dudū primas partes in oi dicendi genere Italia: fel-
sinaꝓ uirum tribuit) in latinū sermonē uertere nō
est dedignatus. Idē Boccatil interpres esse uoluit.
Maximi igit spectatissimꝓ honoris pmiꝓ loco
duximꝓ: q; me lermomis tui uernaculi (quo egregie
polles) interprē esse uoluit. Illud nepe inter preci-
pua humanitatis gēnera locari solet: cū a pcepto-
re discipulus ad honores egregios eleuat: Tātū igit
tur te facio: q;ti Iocrates Platoꝓ Socrate, sanctissi-
mū fecere. Nec cedo ea ire Theophrasto: q; Aristo-
tilē peripatericoꝓ fundatore maximū maxi fecit.
Ecq;d priscos cito: cū satis supꝓ mea i te obserua-
bilꝓ piꝓq; ueneratio, lōge lateꝓ p fines Germaniꝓ
eruditissimāꝓ latī scholas, me preconē, itonuī.
testis ē mihi Hubertinꝓ Clericus Cresserinas: mar-
chionis Montiserrati & pedemōranꝓ poeta stipe-
darius: apud quē Cassale tuā egregiam uirtutem
expolui. Interrogabat enī uir iste apprimē doctus:

Ergebnis Zeilensegmentierung

Evaluation I – Barclay I

- Werk: Barclay's Narrenschiff.
 - Nachdruck von 1509.
 - 570 Seiten.
- Zeitaufwand:
 - LAREX: Komplettes Werk in 2h 18min.
 - Aletheia: In gleicher Zeit nur 160 Seiten.



Erwartete Segmentierung zweier Beispielseiten

Evaluation I – Barclay II

- Erfassung der benötigten Änderungen auf den ersten 200 Seiten.
- Sehr simple Änderungen überwiegen.
- Spezielle Nutzeranforderungen können Aufwand erheblich steigern.

Art der Änderung	#	Aufwand
Irrelevante Region entfernen	161	Rechtsklick
Irrelevantes Bild entfernen	90	Rechtsklick
Typkorrektur: Marginalie → Bildbeschreibung	65	Doppelklick und Typ
Abtrennen der Bogensignatur	31	Linie zeichnen
Typkorrektur: Fließtext → Marginalie	17	Doppelklick und Typ
Verbinden zweier Textblöcke	10	Rechteck und Typ
Manuelles Auszeichnen eines Bildes	2	Rechteck und Typ
Manuelles Auszeichnen der Seitenzahl	2	Rechteck und Typ

Evaluation II – Reale Anwendung auf *Narrenschiffe*

- Im Rahmen des [*Narragonien digital-Projekts*](#).
- Erzielte Ergebnisse:

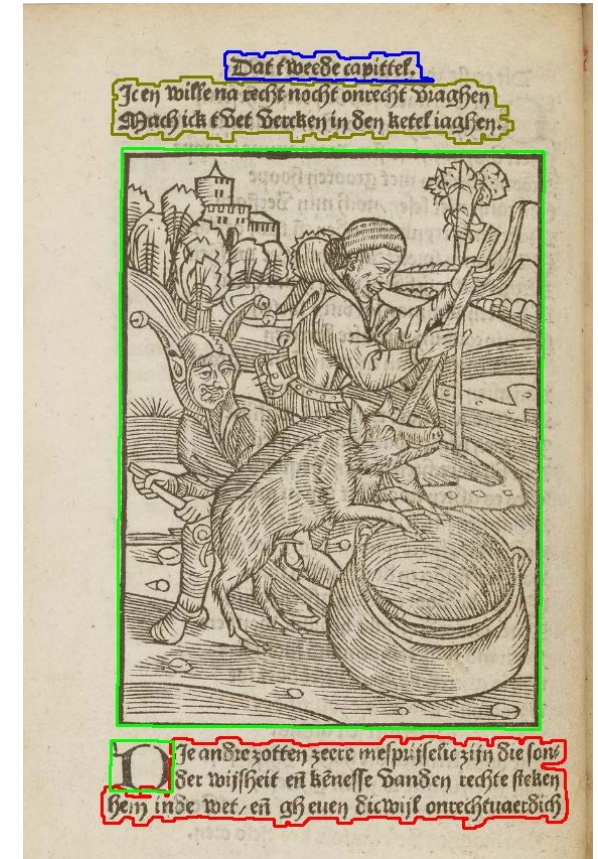
	Barclay	GW5064	GW5066
Anzahl Seiten	570	351	234
DPI	400	400	96
Grobsegmentierung	2h 18min	2h 20min	2h 00min
Feinsegmentierung	2h 40min	2h 15min	1h 45min
Zeichengenauigkeit auf Testset	98.40%	98.13%	98.80%

Folio
 Ast ubi doctrine certamen uoluitur: inq[ui]
 Aedibus in nostris librorum culta supellex
 Eminet: & chartis uiuo contentus operis:
 Quas uideo ignorans: iuuat & me copia sola.
 Constituit quondam diues Prolomeus: haberet
 Vt libros toto quæsitos undiq[ue] mundo.
 Quos grandes rerum thelauros esse putabat:
 Non tamen archan[ge] legis documenta tenebat:
 Quis sine non poterat uite disponere cursum:
 En pariter teneo numerosa uolumina, tardus
 Pauca lego: uiridi contentus tegmine libri.
 Cur uellem studio sensus turbare frequenti?
 Aut tam sollicitis animum confundere rebus:
 Qui studet, assiduo mori, fit stultus et amens.
 Seu studeat: seu nō: dominus tamen esse uocabor
 Et possum studio socum disponere nostro:
 Qui pro me sapiat: doctasq[ue] examiner artes.
 At si cum doctis uerlos: concedere malo
 Omnia: ne cogar fors uerba latina profari.
 Theutonicos inter balbos sum maximus auctor:
 Cum quibus in cassum sparguntur uerba latina.
 O uos doctores: qui grādā nomina fertis:
 Respicite antiquos patres: iurisq[ue] peritos.
 Non in candidulis pensabant dogmata libris:
 Arte sed ingenua sitibundū pectus alebant.
 Auriculis aini regitur sed magna cæterua:
 Ille agit, inq[ui] scobrem trudit ubiq[ue] suū.

De bonis consultoribus
 Ciuilis quicunq[ue] gerit consulta senatus:
 Iustitiamq[ue] uidens, sena aliena probat:
 Condemnatq[ue] graui miseros errore potenter:
 Ille agit, inq[ui] scobrem trudit ubiq[ue] suū.

Prouer. V

Beispielseite GW5064

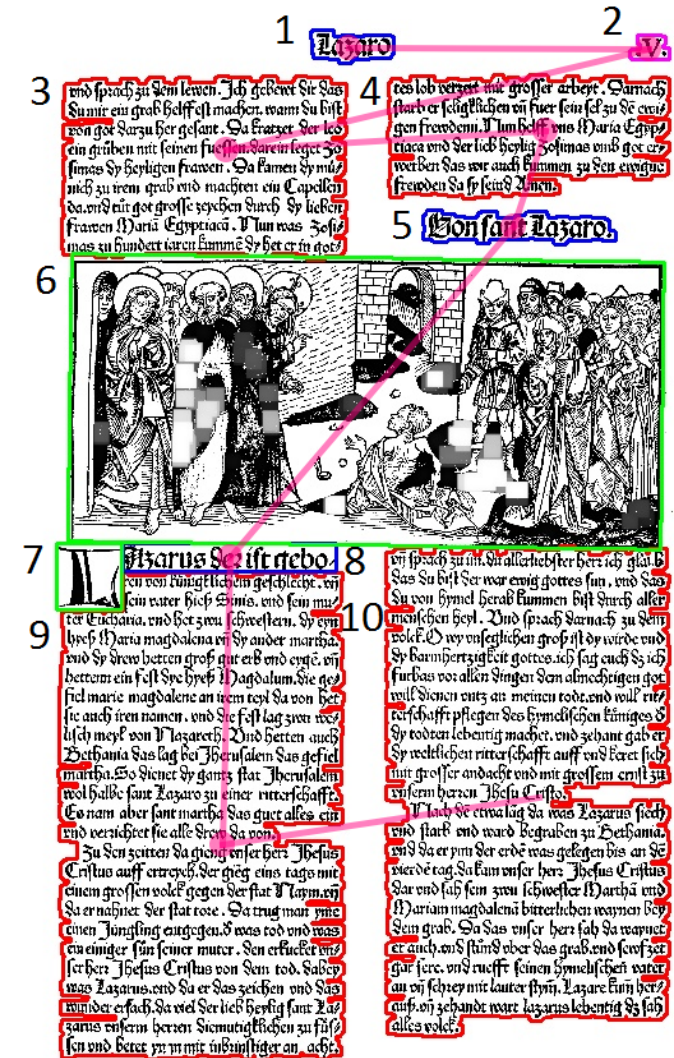


Beispielseite GW5066

Evaluation III – Der Heiligen Leben

- Test einer vollautomatischen Segmentierung.
- Werk: I.t.f. 954 – Der Heiligen Leben (1488).
- Vorherige Anpassung auf Codeebene nötig!
- Vergleich des Zeitaufwands der Segmentierung und des resultierenden OCR-Ergebnisses.
- Ergebnisse:

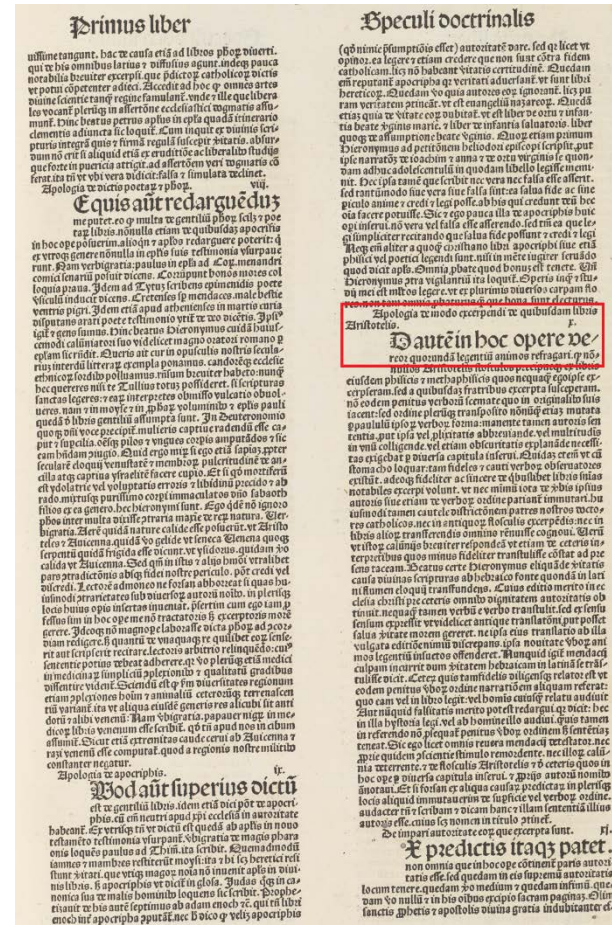
	Aletheia	LAREX
Manueller Zeitaufwand	ca. 100h	ca. 1h
Zeichengenauigkeit (95% KI)	97,57 (20) %	97,35 (28) %
Wortgenauigkeit (95% KI)	92,19 (51) %	91,84 (62) %



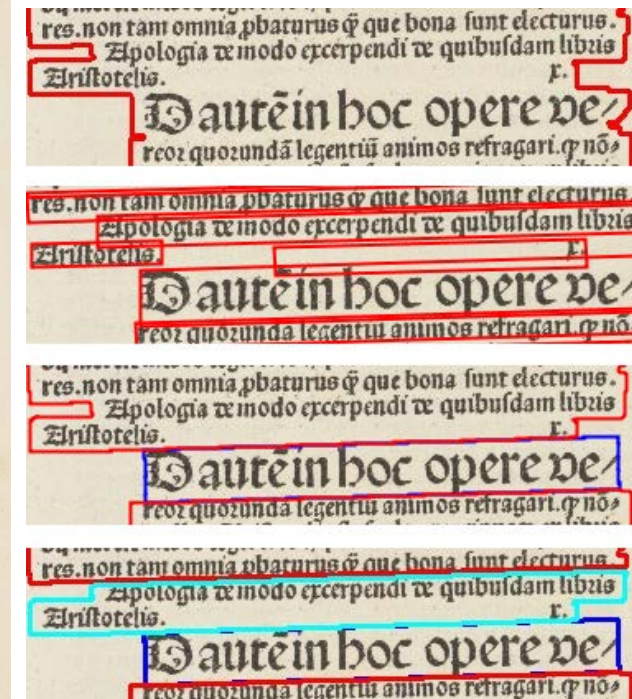
Korrekte Segmentierung einer Beispielseite

Aktuell: Vinzenz von Beauvais' Specula

- „Enzyklopädien“ des Spätmittelalters (Druck 1481-1486).
- Vier umfassende Werke (knapp 3300 Seiten).
- Homogenes, zweispaltiges Layout...
- ... leider mit ein paar Extras.



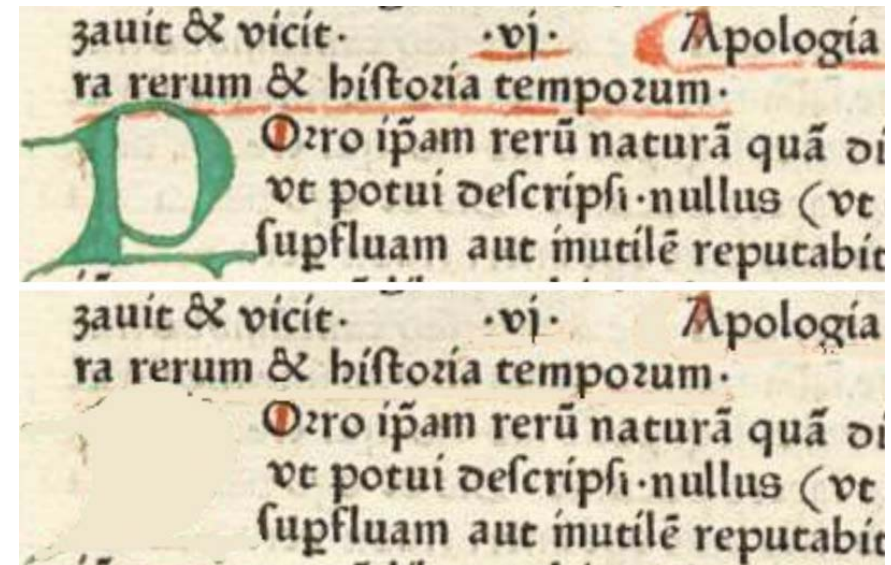
Beispielscan (beschnitten)



Schritte einer möglichen Segmentierung

Diskussion und Ausblick

- Intuitives Tool zur semi-automatischen Segmentierung von frühen Buchdrucken.
- Bereits sehr gute Ergebnisse in realen Anwendungsszenarien.
- Trotzdem noch viele Baustellen:
 - (Web-)Oberfläche.
 - Ausbau des Regelsystems.
 - Adaptation an neuere Drucke.
 - (Optionale,) robustere Bilderkennung.
 - Text in „Bild“.
 - ...
 - Integration Vorverarbeitung.
 - Integration Initialenentfernung.
 - ...



Beispiel Initialenentfernung

Vielen Dank für Ihre Aufmerksamkeit!

- Reul, C., Springmann, U., Puppe, F.: LAREX – A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books.
 - Eingereicht bei [DATECH 2017](#). Verfügbar auf [arXiv.org](#).
 - [Tool Homepage](#) (im Aufbau).
- Reul, C., Dittrich, M., Gruner, M.: Case Study of a highly automated Layout Analysis and OCR of an incunabulum: ‚Der Heiligen Leben‘ (1488).
 - Eingereicht bei [DATECH 2017](#). Verfügbar auf [arXiv.org](#).
 - [Original Scans ‚Der Heiligen Leben‘](#) (Virtuelle Bibliothek Uni Würzburg).
 - [OCR-Ergebnis](#) der 41 Testseiten.